BIT
1998, Vol. 38, No. 4, pp. 674–684

0006-3835/98/3804-0674 $12.00

ON NEWTON'S METHOD FOR HUBER'S ROBUST M-ESTIMATION PROBLEMS IN LINEAR REGRESSION *

B. CHEN[1] and M. Ç. PINAR[2] †

[1] *Department of Management and Systems, Washington State University, Pullman WA 99164-4736, USA. email: chenbi@wsu.edu*

[2] *Department of Industrial Engineering, Bilkent University, Bilkent, Ankara 06533, Turkey. email: mustafap@bilkent.edu.tr*

Abstract.

The Newton method of Madsen and Nielsen (1990) for computing Huber's robust M-estimate in linear regression is considered. The original method was proved to converge finitely for full rank problems under some additional restrictions on the choice of the search direction and the step length in some degenerate cases. It was later observed that these requirements can be relaxed in a practical implementation while preserving the effectiveness and even improving the efficiency of the method. In the present paper these enhancements to the original algorithm are studied and the finite termination property of the algorithm is proved without any assumptions on the M-estimation problems.

AMS subject classification: 62J05, 65D10, 65F20, 65U05.

Key words: Huber's M-estimate, robust regression, Newton's method, finite convergence.

1 Introduction.

In this paper we study a Newton-type method for Huber's robust M-estimator in linear regression. This method was proposed by Madsen and Nielsen in [7]. It was proved to converge finitely for full rank problems through an elegant analysis that delineated some essential features of the algorithm. The algorithm is known to be quite efficient as reported in [7]. It was later used successfully as a subroutine for linear ℓ_1 estimation [8] and for linear programming [9]. Interestingly, it was observed in [10, 11] that those features of the algorithm essential for the finite convergence analysis could be removed in an implementation without affecting effectiveness and efficiency of the algorithm. In this regard, a question arose whether the algorithm retains its finite convergence under these modifications. This discrepancy between theory and practice remained unexplained thus far, to the best of our knowledge. On the other hand, it seems difficult to modify the analysis of Madsen and Nielsen to cover these enhancements. The purpose

*Received January 1997. Revised November 1997. Communicated by Kaj Madsen.
†Research supported by NATO Collaborative Research Grant CRG-94-0609.

of the present paper is to bridge this gap between theory and practice for this important algorithm. To this end, we give a modified version of the algorithm and provide a new finite convergence analysis. While it is possible in practice to reduce a rank deficient problem to a full rank one using a preprocessor prior to the execution of the algorithm, our analysis shows that the full rank assumption is not necessary for the finite convergence property to hold.

Robust estimation is concerned with identifying "outliers" among data points and giving them less weight. Huber' s M-estimator is essentially the least squares estimator, which uses the ℓ_1-norm for points that are considered outliers with respect to a certain threshold. Hence, the Huber criterion is less sensitive to the presence of outliers [3].

Let $A \in \Re^{m\times n}$, $b \in \Re^m$. Denote $A^T = [a_1 a_2 \dots a_m]$. We are interested in minimizing a residual vector $r(x) = Ax - b$ with

$$r_i(x) = a_i^T x - b_i, \quad i = 1, \dots, m,$$

using the Huber function

$$\rho(t) = \begin{cases} \frac{1}{2\gamma}t^2 & \text{if } |t| < \gamma \\ |t| - \frac{1}{2}\gamma & \text{if } |t| \geq \gamma \end{cases} \tag{1.1}$$

with a tuning constant $\gamma > 0$. Huber's M-estimate is a minimizer $x^* \in \Re^n$ of the function

$$F(x) = \sum_{i=1}^{m} \rho(r_i(x)). \tag{1.2}$$

To view this minimization problem in a different format, we introduce the following index sets at a point $x \in \Re^n$:

$$\begin{aligned} \mathcal{I}_-(x) &= \{i | r_i(x) < -\gamma\}, & \bar{\mathcal{I}}_-(x) &= \{i | r_i(x) \leq -\gamma\}, \\ \mathcal{I}(x) &= \{i | |r_i(x)| \leq \gamma\}, & \bar{\mathcal{I}}(x) &= \{i | |r_i(x)| < \gamma\}, \\ \mathcal{I}_+(x) &= \{i | r_i(x) > \gamma\}, & \bar{\mathcal{I}}_+(x) &= \{i | r_i(x) \geq \gamma\}. \end{aligned}$$

We call $\mathcal{I}(x)$ and $\bar{\mathcal{I}}(x)$ the active set and the strictly active set at x, respectively. In addition, we introduce the following "sign vector" $s \in \Re^m$ associated with the above index sets:

$$s(x) = [s_1(x) s_2(x) \dots s_m(x)]^T$$

with

$$s_i(x) = \begin{cases} -1 & \text{if } i \in \mathcal{I}_-(x) \\ 0 & \text{if } i \in \mathcal{I}(x) \\ 1 & \text{if } i \in \mathcal{I}_+(x). \end{cases}$$

A sign vector s is feasible if there exists an $x \in R^n$ such that $s = s(x)$. We also define the following diagonal matrix $W \in \Re^{m\times m}$ associated with s:

$$W(x) = \operatorname{diag}[w_1(x), w_2(x), \dots, w_m(x)], \tag{1.3}$$

where

$$w_i(x) = 1 - s_i^2(x). \tag{1.4}$$

Clearly, $w_i(x) = 1$ for all $i \in \mathcal{I}(x)$ and $w_i(x) = 0$ otherwise. Similarly, $\bar{W}$ and $\bar{s}$ can be defined based on index sets $\bar{\mathcal{I}}_-$, $\bar{\mathcal{I}}$, and $\bar{\mathcal{I}}_+$.

Using the above notation, Huber's M-estimation problem can be expressed as the following minimization problem:

[**P**]

$$\text{minimize } F(x) = \frac{1}{2\gamma} r^T(x) W(x) r(x) + s^T(x) \left[r(x) - \frac{\gamma}{2} s(x) \right]. \tag{1.5}$$

The following properties of F have been shown in [7]:

LEMMA 1.1. *The following properties hold for F defined in* (1.5)*:*

1. *F is piecewise quadratic, convex, and once differentiable at those x such that $|r_i(x)| = \gamma$ for some $i = 1, \ldots, m$.*
2. *F is bounded below and therefore has a finite minimizer.*

The gradient of F is given by

$$F'(x) = A^T \left[\frac{1}{\gamma} W(x) r(x) + s(x) \right]. \tag{1.6}$$

Let X be the set of all solutions of **P**. Clearly, $x \in X$ if and only if $F'(x) = 0$. In addition, Madsen and Nielsen [8] have shown the following properties about the solution set X:

LEMMA 1.2. *Both $\bar{s}(x)$ and $r_i(x)$, with $i \in \bar{\mathcal{I}}(x)$, are constant for all $x \in X$.*

Let s be a feasible sign vector, I and W be the corresponding active set and diagonal matrix, respectively. Define $\mathcal{C}_s = cl\{x | s(x) = s\}$ as a set of x induced by s. Clearly, $F(x)$ is identical to the following quadratic function $F_s(x)$ on the set $\mathcal{C}_s$:

$$F_s(x) = \frac{1}{2\gamma} r(x)^T W r(x) + s^T \left[r(x) - \frac{\gamma}{2} s \right].$$

2 The Newton method of Madsen and Nielsen.

The Newton method of Madsen and Nielsen [7] is a modified Newton method with a line search procedure. We will refer to this algorithm as the MN algorithm for convenience.

In light of the above discussion, the MN algorithm consists of inspecting the domains $\mathcal{C}_s$ to find the quadratic representation of F where the global minimizer is located. A search direction h is computed by minimizing the quadratic $F_s(x)$ where s is the sign vector of the current iterate. More precisely, let x be the current iterate and $s = s(x)$ and $W = W(x)$, the MN algorithm uses the following system of equations to generate a search direction:

$$F_s'' h = -F'(x). \tag{2.1}$$

This system can be expressed as

$$(A^T W A)h = -A^T[Wr(x) + \gamma s]. \tag{2.2}$$

Clearly, $x + h$ minimizes the quadratic F_s for any h that solves (2.2). If in addition $x + h \in \mathcal{C}_s$, then $F'(x + h) = F_s'(x + h) = 0$ and $x + h$ is also a minimizer of F.

We now present the MN algorithm as described in [7] for comparison purposes:

stop = false
repeat
  $s = s(x)$
  if F_s'' is positive definite **then**
    find h as the unique solution to (2.2)
    if $x + h \in \mathcal{C}_s$ **then**
      $x \leftarrow x + h$
      stop = true
    else
      $x \leftarrow x + \lambda h$ (line search)
    endif
  elseif F_s'' is positive semi-definite and (2.2) is consistent
    find h as the minimum norm solution of (2.2)
    if $x + h \in \mathcal{C}_s$ **then**
      $x \leftarrow x + h$
      stop = true
    else
      $x \leftarrow x + \alpha_1 h$
    endif
  else (F_s'' is positive semi-definite and (2.2) is inconsistent)
    find h as the solution of $Dh = -F_s'(x)$ for some
    positive definite matrix D.
    $x \leftarrow x + \lambda h$ (line search)
  endif
until stop.

REMARK 2.1. The case where F_s'' is positive semi-definite and the system (2.2) is consistent is called the *degenerate* case in [7]. In this case, the MN algorithm requires the minimum norm solution h of (2.2), which is in general computationally expensive. Furthermore, if $x + h$ is not a minimizer of F, the algorithm then proceeds with a restrictive line search; the next iterate is found by moving to the first breakpoint α_1 along h, i.e., the smallest value of α where $s(x + \alpha) \neq s(x)$.

REMARK 2.2. Regarding the choice of the positive definite matrix D in the algorithm, it is required that the smallest eigenvalue of D be uniformly bounded below by a positive constant. For example, one can choose D as the identity matrix, or as $A^T W A + \epsilon I$ with a positive number ϵ.

Madsen and Nielsen [7] showed that their algorithm converges finitely.

THEOREM 2.1. *If A has full rank, the above algorithm stops at a minimizer after a finite number of iterations.*

3 A modified Newton algorithm.

We consider the following enhancements to the MN algorithm.

1. When the system (2.2) has multiple solutions, our modified algorithm does not restrict the search direction to be the minimum norm solution of (2.2). This allows us to use a basic solution as in the implementation by Nielsen [10, 11].

2. We carry out a line search regardless of whether the system (2.2) is consistent or not (this is used in the implementations of [8, 9]). However, the original MN algorithm restricts the step length when the system has multiple solutions.

3. We establish the finite convergence without the assumption that A has full rank.

The modified algorithm can be stated as follows.

```
stop = false
repeat
    s = s(x)
    if (2.2) is consistent then
            find h as any solution to (2.2) such that ||h|| ≤ κ||h_m||
            (cf. Remark 3.1)
            if x + h ∈ C_s then
                    x ← x + h
                    stop = true, x is a solution of P
            endif
    else
            find h as the solution of Dh = −F_s'(x) (cf. Remark 2.2)
    endif
    x ← x + λ̄h (line search, cf. Remarks 3.2 and 3.3)
until stop.
```

REMARK 3.1. In case the system (2.2) is consistent, h_m is used to denote the minimum norm solution of (2.2), and $\kappa \geq 1$ is a constant. It is well known that any basic solution h_b to (2.2) computed from a QR decomposition with column pivoting of A^TWA satisfies $\|h_b\| \leq \kappa \|h_m\|$ for some constant $\kappa \geq 1$; see p. 244 of [2] for details.

REMARK 3.2. As indicated by one of the referees, the above modified Newton algorithm is closely related to a Newton algorithm by Li and Swetits [6] for solving strictly convex quadratic programs. The major conceptual difference lies in the choice of the step size in the line search phase. While the Li–Swetits

algorithm restricts the step size to be less than or equal to 1, our algorithm removes this restriction.

REMARK 3.3. The step size $\bar{\lambda}$ in the modified Newton algorithm is not uniquely determined in general. We choose $\bar{\lambda}$ as the smallest minimizer of F in the direction h when there are multiple solutions to the exact line search.

The following result states that h as determined by the above algorithm is a strictly descent direction of F at x.

LEMMA 3.1. *Let $\{x^k\}$ be any sequence generated by the modified Newton method. Let h^k be the search direction at x^k with $s^k = s(x^k)$ and $W^k = W(x^k)$. Then*

$$(h^k)^T F'(x^k) \leq -c\|h^k\|^2 \tag{3.1}$$

for some constant $c > 0$. Furthermore, $(h^k)^T F'(x^k) = 0$ only if $F'(x^k) = 0$.

PROOF. The result clearly holds if h^k is generated from $D^k h = -F'_s(x^k)$ since the smallest eigenvalue of D^k is uniformly bounded below by a positive constant by construction.

We next show that the result also holds if h^k is a solution of (2.2). Since $A^T W^k A$ is a symmetric positive semidefinite matrix, it has a set of orthonormal eigenvectors. Let α_j^k and e_j^k denote the eigenvalues and eigenvectors of $A^T W^k A$. Assume, without loss of generality, that the first p^k eigenvalues are positive and have been arranged in non-increasing order, and the remaining eigenvalues are equal to zero. Since equation (2.2) is consistent, the expansion of $F'(x^k)$ with respect to the eigenvectors should have the following form:

$$F'(x^k) = \sum_{j=1}^{p^k} \beta_j^k e_j^k,$$

for some β_j^k, $j = 1, \ldots, p^k$. As a result, a general solution of equation (2.2) is given by

$$h^k = -\sum_{j=1}^{p^k} \frac{\beta_j^k}{\alpha_j^k} e_j^k - \sum_{j=p^k+1}^{n} \xi_j^k e_j^k,$$

for some ξ_j^k, $j = p^k + 1, \ldots, n$, while the minimum norm solution h_m^k can be expressed as

$$h_m^k = -\sum_{j=1}^{p^k} \frac{\beta_j^k}{\alpha_j^k} e_j^k,$$

It follows that for all k we have

$$(h^k)^T F'(x^k) = -\sum_{j=1}^{p^k} \frac{(\beta_j^k)^2}{\alpha_j^k} = -\sum_{j=1}^{p^k} \left(\frac{\beta_j^k}{\alpha_j^k}\right)^2 \alpha_j^k \leq -\alpha_{p^k}\|h_m^k\|^2 \leq -c'\|h_m^k\|^2,$$

where the existence of the constant $c' > 0$ in the last inequality follows from the fact that there is only a finite number of matrices $A^T W^k A$. Since $\|h^k\| \leq \kappa\|h_m^k\|$,

we have

$$(h^k)^T F'(x^k) \leq -c\|h^k\|^2 \quad \forall k.$$

with $c = c'/\kappa^2$. In addition, if $(h^k)^T F'(x^k) = 0$, then $\|h^k\| = 0$, which implies, by equation (2.2), that $F'(x^k) = 0$. □

Let h be the descent direction generated at x by the algorithm. Unless the system (2.2) is consistent and $x + h \in C_s$, in which case the algorithm stops, the algorithm proceeds with a line search of F along direction h. More precisely, the line search procedure looks for the smallest step length that minimizes the function

$$\theta(\lambda) = F(x + \lambda h).$$

Clearly, θ is a univariate, once differentiable, convex, and piecewise quadratic function. Moreover, it is bounded below since F is bounded below. Therefore, θ has a finite minimizer $\bar{\lambda}$ such that $\theta'(\bar{\lambda}) = 0$. In addition, $\bar{\lambda} > 0$ since h is a descent direction and $\theta'(0) < 0$. Let

$$\theta_s(\lambda) = F_s(x + \lambda h).$$

The following result is obvious since F_s is a convex quadratic function and $x + h$ is a minimizer of F_s. We will need the result later for the finite convergence proof.

LEMMA 3.2. *Let x be any point such that $F'(x) \neq 0$ and $s = s(x)$. Suppose equation (2.2) is consistent and h is a solution of (2.2). Then $\theta_s'(1) = 0$ and $\theta_s'(\lambda) < 0$ for all $0 \leq \lambda < 1$.*

To locate $\bar{\lambda}$, we search for a zero of the non-decreasing piecewise linear smooth function θ'. Let $\mathcal{K} = \{\lambda_k\}$ be the set of positive kink points of θ'. Clearly, $|\mathcal{K}| \leq 2m$. For simplicity, assume that all kink points $\lambda_k \in \mathcal{K}$ are sorted in ascending order. Then the zero of θ' should be in the interval such that $\theta'(\lambda_{j-1}) < 0$ and $\theta'(\lambda_j) \geq 0$. Once the interval is identified, the zero of θ' can be efficiently and accurately calculated since θ' is a linear function over the interval. Furthermore, the quantity $\theta'(\lambda_j)$ can be easily updated from $\theta'(\lambda_{j-1})$ since the move from λ_{j-1} to λ_j only affects one term in the defining equation of θ'. Issues related to an efficient implementation of the line search is discussed in detail in [7, 11]. It was also pointed out by one of the referees that the line search can be performed in $O(m)$ flops using an algorithm by Pardalos and Kovoor [12] for singly constrained quadratic programs.

Since h generated by the modified algorithm is a strict descent direction of F by Lemma 3.1, the step length $\bar{\lambda}$ is obtained by the exact line search, and F is bounded below, we have the following global convergence result for the algorithm:

THEOREM 3.3. *Let $\{x^k\}$ be a sequence generated by the above algorithm. Then either $F'(x^k) = 0$ for some k or $F'(x^k) \to 0$.*

PROOF. Suppose $F'(x^k) \neq 0$ for all k. Let h^k be the strictly descent direction generated at x^k and $\bar{\lambda}^k > 0$ be the corresponding step length obtained by the

exact line search. Since F is bounded below, by the standard step length analysis (see for example the proof of Theorem 6.3.3 of [1]), we have

$$\lim_{k\to\infty} \frac{F'(x^k)^T h^k}{\|h^k\|} = 0. \tag{3.2}$$

By Lemma 3.1,

$$(h^k)^T F'(x^k) \leq -c\|h^k\|^2$$

holds for all k and some constant $c > 0$. This implies that $\|h^k\| \to 0$, and therefore, $F'(x^k) \to 0$. □

By a slight abuse of notation, let $F(X)$ denote the minimum value of F. Since F is convex and the modified algorithm is descent, Theorem 3.3 together with Lemma 3.3 of Li and Swetits [6] imply that $F(x^k)$ converges from above to $F(X)$.

4 Finite termination.

In this section we will show that the modified Newton's algorithm terminates with a minimizer of F in a finite number of iterations.

For any $\epsilon \geq F(X)$, define the following level set for F:

$$L(\epsilon) = \{x \in \Re^n | F(x) \leq \epsilon\}.$$

Since F is a convex function, the level set $L(\epsilon)$ is also convex. In addition, $L(\epsilon) \supseteq X$ for all $\epsilon \geq F(X)$ and $L(\epsilon) \to X$ as ϵ approaches $F(X)$ by Corollary 2.8 of [4]. The next result shows that all points in a level set $L(\epsilon)$ with ϵ sufficiently close to $F(X)$ will have similar index sets, with the difference only in the active set.

LEMMA 4.1. *There exists an $\epsilon_1 > F(X)$ such that $\mathcal{I}_-(x_1) \cap \mathcal{I}_+(x_2) = \emptyset$ for all $x_1, x_2 \in L(\epsilon_1)$. If in addition $\mathcal{I}(x_1) = \mathcal{I}(x_2)$ then $s(x_1) = s(x_2)$.*

PROOF. By Lemma 1.1, both $\bar{s}(x^*)$ and $|r_i(x^*)| < \gamma$, $i \in \bar{\mathcal{I}}(x^*)$, are constant for all $x^* \in X$. Since $r(x)$ is continuous in x, there exists a $\delta > 0$ such that for any $x^* \in X$ and any x satisfying $||x - x^*|| < \delta$, we have

$$\mathcal{I}_-(x) \subseteq \bar{\mathcal{I}}_-(x^*) = \bar{\mathcal{I}}_-(X) \quad \text{and} \quad \mathcal{I}_+(x) \subseteq \bar{\mathcal{I}}_+(x^*) = \bar{\mathcal{I}}_+(X).$$

Since $\bar{\mathcal{I}}_-(X) \cap \bar{\mathcal{I}}_+(X) = \emptyset$, it follows that there exists an open neighborhood $N \supset X$ such that $\mathcal{I}_-(x_1) \cap \mathcal{I}_+(x_2) = \emptyset$ for all $x_1, x_2 \in N$. The first result then follows from the fact that $L(\epsilon) \supseteq X$ for all $\epsilon \geq F(X)$ and $L(\epsilon) \to X$ as ϵ approaches $F(X)$. The second result is an immediate consequence of the first result. □

Denote by X_s the set of all minimizers of F_s, if it exists. The next result shows that the quadratic function F_s induced by any point $x \in L(\epsilon)$ with ϵ sufficiently close to $F(X)$ will have a minimizer in X. As pointed out by a referee, a more general form of this result was given in Lemma 3.7 of [6].

LEMMA 4.2. *There exists an $\epsilon_2 > F(X)$ such that $X \cap X_s \neq \emptyset$ for any $x \in L(\epsilon_2)$ and $s = s(x)$.*

PROOF. Let $\delta_1 > F(X)$ be arbitrary. There is only a finite number of sign matrices, say W_l, $l = 1, \ldots, l_2$, such that the corresponding Q-subset $\mathcal{C}_{s_l}$ intersects with the level set $L(\delta_1)$. Define

$$\epsilon_l = \min_{x \in \mathcal{C}_{s_l} \cap L(\delta_1)} F(x), l = 1, \ldots, l_2.$$

Clearly, $\epsilon_l \geq F(X)$ for all $l = 1, \ldots, l_2$. If $\epsilon_l = F(x^*) = F(X)$ for some $x^* \in \mathcal{C}_{s_l} \cap L(\delta_1)$, then $x^* \in X$. In addition, $x^* \in X_{s_l}$ since $x^* \in \mathcal{C}_{s_l}$ and $F'_{s_l}(x^*) = F'(x^*) = 0$. Therefore, if $\epsilon_l = F(X)$ for all $l = 1, \ldots, l_2$, we may choose $\epsilon_2 = \delta_1$ and the result is proved. Otherwise, let δ_2 be the smallest ϵ_l among all $l = 1, \ldots, l_2$ such that $\epsilon_l > F(X)$. Clearly, $\delta_2 > F(X)$. The result is proved by choosing any ϵ_2 such that $\delta_2 > \epsilon_2 > F(X)$. □

Notice, however, that the above result did not claim that $X_s \subseteq X$ for $s = s(x)$ and $x \in L(\epsilon_2)$. Indeed, if F_s has multiple minimizers, it could happen that some minimizers of F_s belong to X and others not. The following result studies the properties of the minimizers of F_s.

LEMMA 4.3. *Let $x \in \Re^n$, $s = s(x)$, and $W = W(x)$. Then we have the following:*

1. *If $x_1 \in X_s$ and $x_2 \in X_s$, then $x_1 - x_2 \in \mathcal{N}(WA)$, where $\mathcal{N}(C)$ represents the null space of matrix C.*
2. *If there exists an $x_1 \in X_s$ such that $x_1 \in \mathcal{C}_s$, then $\mathcal{I}(x^*) \supseteq \mathcal{I}(x)$ for all $x^* \in X_s$.*

PROOF. Since x_1 and $x_2 \in X_s$, both $x_1 - x$ and $x_2 - x$ are solutions of (2.2). It follows that $(A^T W A)(x_1 - x_2) = 0$. Part 1 then follows from the fact that $\mathcal{N}(A^T W A) = \mathcal{N}(WA)$. For part 2, it suffices to show that $|r_i(x^*)| \leq \gamma$ for all $i \in \mathcal{I}(x)$. Indeed, by part 1, we have $(x_1 - x^*)^T a_i = 0$ for all $i \in \mathcal{I}(x)$. Therefore, $|r_i(x^*)| = |r_i(x_1)| \leq \gamma$ for all $i \in \mathcal{I}(x)$. □

Now we are ready to show the finite convergence for the modified MN algorithm.

THEOREM 4.4. *The modified Newton algorithm finds a minimizer of F in a finite number of iterations.*

PROOF. Let $\{x^k\}$ be a sequence generated by the modified Newton's algorithm such that $F'(x^k) \neq 0$ for all k. Set $\epsilon = \min\{\epsilon_1, \epsilon_2\}$, where ϵ_1 and ϵ_2 are defined in Lemma 4.1 and Lemma 4.2, respectively. Since $F(x^k)$ converges from above to $F(X)$, there exists an integer $K > 0$ such that $F(x^k) \leq \epsilon$ and $x^k \in L(\epsilon)$ for all $k \geq K$. Let $k \geq K$, $s^k = s(x^k)$, and $W^k = W(x^k)$. By Lemma 4.2, F_s has a minimizer in X and therefore, (2.2) is consistent. Thus the next iterate generated by the algorithm is $x^{k+1} = x^k + \bar{\lambda}^k h^k$, where h^k is a solution of (2.2) and $\bar{\lambda}^k$ is determined by the exact line search. We claim that $\bar{\lambda}^k \leq 1$. Suppose on the contrary that $\bar{\lambda}^k > 1$. Then $F(x^k + h^k) < F(x^k)$ and $\theta'(1) < 0$ since h^k is a strictly descent direction of F at x^k by Lemma 3.1. By Lemma 4.3, $\mathcal{I}(x^k + h^k) \supseteq \mathcal{I}(x^k)$. It follows that

$$\mathcal{I}(x^k + \lambda h^k) \supseteq \mathcal{I}(x^k) \quad \forall \lambda \in [0, 1], \tag{4.1}$$

since

$$\begin{aligned} |r_i(x^k + \lambda h^k)| &= |\lambda r_i(x^k) + (1-\lambda) r_i(x^k + h^k)| \\ &\leq \lambda |r_i(x^k)| + (1-\lambda)|r_i(x^k + h^k)| \\ &\leq \gamma \end{aligned}$$

for all $i \in \mathcal{I}(x^k)$. Define

$$\delta(\lambda) = \theta(\lambda) - \theta_{s^k}(\lambda) \quad \text{for } \lambda \in [0,1].$$

In view of the definition of the active set at x^k, we have

$$\delta(\lambda) = \sum_{i \notin I(x^k)} \rho(r_i(x^k + \lambda h^k)) - \sum_{i \notin I(x^k)} s_i^k \left[r_i(x^k + \lambda h^k) - \frac{\gamma}{2} s_i^k \right].$$

Since the first summation in the expression of δ is a convex function of λ, and the second summation is a linear function of λ, $\delta(\lambda)$ is a continuously differentiable convex function for $\lambda \in [0,1]$. Since

$$\delta'(0) = \theta'(0) - \theta'_{s^k}(0) = 0,$$

we have $\delta'(1) \geq 0$. Hence, $\theta'_{s^k}(1) \leq \theta'(1) < 0$. However, this contradicts the fact that h^k is a solution of (2.2) and thus $\theta'_{s^k}(1) = 0$. Therefore, $\bar{\lambda}^k \leq 1$. Using (4.1) again, we have

$$\mathcal{I}(x^{k+1}) = \mathcal{I}(x^k + \bar{\lambda}^k h^k) \supseteq \mathcal{I}(x^k).$$

In addition, $x^{k+1} \in L(\epsilon)$ since $F(x^{k+1}) < F(x^k)$. Suppose $\mathcal{I}(x^{k+1}) = \mathcal{I}(x^k)$. By Lemma 4.1, we have $s(x^{k+1}) = s(x^k)$. Thus, $\theta'_{s^k}(\bar{\lambda}^k) = \theta'(\bar{\lambda}^k) = 0$. By Lemma 3.2, $\bar{\lambda}^k = 1$. Therefore, $x^{k+1} = x^k + h^k$ and $x^k + h^k \in \mathcal{C}_{s^k}$. It follows that x^{k+1} is a minimizer of F since $F'(x^k + h^k) = F'_{s^k}(x^k + h^k) = 0$. In summary, we have shown that either x^{k+1} is a minimizer of F and the algorithm stops, or $\mathcal{I}(x^{k+1}) \supset \mathcal{I}(x^k)$ and the active set expands. Since $x^{k+1} \in L(\epsilon)$, the above argument can be repeated with x^k replaced by x^{k+1}. However, the active set has only finite cardinality. Therefore, the algorithm must terminate in a finite number of iterations with a minimizer of F. □

Finally, it was brought to our attention by a referee that it is possible to find a Huber M-estimate in $O(m)$ arithmetic operations when n is fixed; see [5] for details. This reference also describes other numerical algorithms for Huber's M-estimate, including a Gauss–Seidel method, matrix splitting methods, and a conjugate gradient method.

Acknowledgments.

The authors are grateful to two anonymous referees for pointing out some gaps in proofs in the first version of the paper and for suggestions that led to improvements of the paper.

REFERENCES

1. J. E. Dennis, Jr. and R. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, SIAM, Philadelphia, PA, 1996.
2. G. H. Golub and C. F. Van Loan, *Matrix Computations*, 2nd ed., Johns Hopkins University Press, Baltimore, MD, 1989.
3. P. Huber, *Robust Statistics*, Wiley, New York, 1981.
4. W. Li, *Error bounds for piecewise convex quadratic programs and applications*, SIAM J. Control Optim., 33 (1995), pp. 1510–1529.
5. W. Li, *Numerical algorithms for the Huber M-estimator problem*, in Approximation Theory VIII, Vol. 1: Approximation and Interpolation, C.K. Chui and L.L. Schumaker, eds., World Scientific Publishing, New York, 1995, pp. 325–334.
6. W. Li and J. Swetits, *A new algorithm for solving strictly convex quadratic programs*, SIAM J. Optim., 7 (1997), pp. 595–619.
7. K. Madsen and H. B. Nielsen, *Finite algorithms for robust linear regression*, BIT, 30 (1990), pp. 682–699.
8. K. Madsen and H. B. Nielsen, *A finite smoothing algorithm for linear ℓ_1 estimation*, SIAM J. Optim., 3 (1993), pp. 68–80.
9. K. Madsen, H. B. Nielsen, and M. Ç. Pınar, *A new finite continuation algorithm for linear programming*, SIAM J. Optim., 6 (1996), pp. 600–616.
10. H. B. Nielsen, *AAFAC: A package of Fortran 77 subprograms for solving $A^T Ax = c$*, Report NI 90–01, Institute for Numerical Analysis, Technical University of Denmark, Lyngby, 1990.
11. H. B. Nielsen, *Implementation of a finite algorithm for linear ℓ_1 estimation*, Report NI 91–01, Institute for Numerical Analysis, Technical University of Denmark, Lyngby, 1991.
12. P. M. Pardalos and N. Kovoor, *An algorithm for a singly constrained class of quadratic programs subject to upper and lower bounds*, Math. Prog., 46 (1990), pp. 321-328.